

Molecular Identifier Visual Prompting and Verifiable Reinforcement Learning for Chemical Reaction Diagram Parsing

Jiahe Song^{1,2*}, Chuang Wang^{2,3*}, Yinfan Wang^{2*}, Hao Zheng^{5*}, Bowen Jiang^{2,4*}, Rui Nie^{2,3*}, Xingjian Wei², Junyuan Gao², Yubin Wang², Bin Wang², Lijun Wu², Jiang Wu^{2†‡}, Qian Yu^{3†}, and Conghui He^{2†}

¹ Shanghai Jiao Tong University, Shanghai, China

² Shanghai Artificial Intelligence Laboratory, Shanghai, China

³ Beihang University, Beijing, China

⁴ Peking University, Beijing, China

⁵ South China Normal University, Guangzhou, China

{wujiang,heconghui}@pjlab.org.cn

Abstract. Reaction diagram parsing (RxnDP) is critical for extracting chemical synthesis information from literature. Although recent Vision-Language Models (VLMs) have emerged as a promising paradigm to automate this complex visual reasoning task, their application is fundamentally bottlenecked by the inability to align visual chemical entities with pre-trained knowledge, alongside the inherent discrepancy between token-level training and reaction-level evaluation. To address these dual challenges, this work enhances VLM-based RxnDP from two complementary perspectives: prompting representation and learning paradigms. First, we propose **Identifier as Visual Prompting (IdtVP)**, which leverages naturally occurring molecule identifiers (e.g., bold numerals like **1a**) to activate the chemical knowledge acquired during VLM pre-training. IdtVP enables powerful zero-shot and out-of-distribution capabilities, outperforming existing prompting strategies. Second, to further optimize performance within fine-tuning paradigms, we introduce **Re³-DAPO**, a reinforcement learning algorithm that leverages verifiable rewards to directly optimize reaction-level metrics, thereby achieving consistent gains over standard supervised fine-tuning. Additionally, we release the **ScannedRxn** benchmark, comprising scanned historical reaction diagrams with real-world artifacts, to rigorously assess model robustness and out-of-distribution ability. Our contributions advance the accuracy and generalization of VLM-based reaction diagram parsing.

Keywords: Reaction Diagram Parsing · Visual Prompt · Reinforcement Learning · Vision Language Model

* Equal contribution. † Corresponding author. ‡ Project lead. Code: <https://github.com/opedatalab/RxnID>. Homepage: <http://chuangwang123.github.io/RxnID/>

1 Introduction

High-quality chemical reaction datasets form the cornerstone of AI for Science research [4], driving the development of automated reaction information extraction from chemical literature [5, 19, 29]. Although existing pipelines heavily rely on text mining, complete synthesis routes and complex mechanisms are primarily communicated through graphical diagrams. As a result, Reaction Diagram Parsing (RxnDP) has emerged as an indispensable task to capture this visually encoded knowledge. Because these diagrams encapsulate rich topological and stereochemical details often omitted in plain text, RxnDP is crucial for unlocking the massive, high-quality data trapped in unstructured bitmaps—a prerequisite for training comprehensive chemical foundation models. Furthermore, visual parsing provides unique structural insights that enable cross-validation with text-based data (as detailed in §Sec. 4.5), ultimately facilitating the construction of highly accurate chemical databases.

While traditional deep learning methods have been widely applied to RxnDP, Vision-Language Models (VLMs) [2, 7, 12, 16] have demonstrated immense potential. Benefiting from breakthroughs in visual and multimodal reasoning, VLMs surpass traditional deep learning approaches [9, 18, 25, 28, 30]. This superiority stems from two advancements: first, the extensive world knowledge acquired through large-scale interleaved image-text pre-training [14, 15, 31]; second, the significant enhancement of cross-domain generalization achieved through Reinforcement Learning [8, 10, 11, 21, 24, 26, 27] from Verifiable Rewards (RLVR) [6, 13, 20, 23] during post-training, moving beyond standard fine-tuning.

Despite these significant advancements in pre-training and RLVR, their specific application to RxnDP remains under-explored. This study aims to address **two core questions**: (1) *How can we better leverage the pre-trained core capabilities of VLMs for chemical diagram understanding?* and (2) *How can we optimize RLVR to improve the performance and generalization of RxnDP?*

Regarding the first challenge, prior work [22] indicated that the conventional **BROS** (**B**box and **R**ole in **O**ne **S**tep) approach is suboptimal for VLMs. To address this, they proposed **BIVP** (**B**ox **I**ndexing as **V**isual **P**rompts), which detects molecules and overlays indexed bounding boxes onto the image as visual prompt. However, this practice introduces a distribution shift from the raw images used in VLM pre-training. We observe that chemical literature frequently uses bold numerals as *molecule identifiers* to annotate reaction diagrams, and the main text commonly employs similar shorthand. Because VLM pre-training corpora contain vast amounts of such interleaved image-text data, models are inherently accustomed to describing chemical reactions using these reference numbers. Capitalizing on this, we propose **IdtVP** (**I**dentifier as **V**isual **P**rompting), which utilizes the inherent identifiers in reaction diagrams as descriptive anchors. Experiments demonstrate that IdtVP significantly outperforms BROS and BIVP in both zero-shot and fine-tuning settings. For instance, on Gemini 3.0 Pro [7], our approach achieves state-of-the-art zero-shot Soft Match F1 scores, proving that IdtVP unlocks the chemical knowledge embedded in VLM pre-training. Fur-

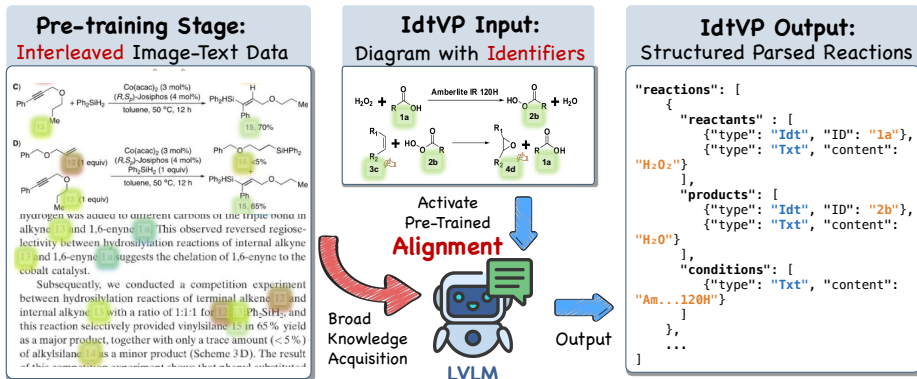


Fig. 1: Overview of the Identifier as Visual Prompting (IdtVP) strategy. **Left:** VLMs learn visual-text alignment from identifier-rich interleaved literature during pre-training. **Middle:** IdtVP inputs diagrams annotated with molecule identifiers to activate this pre-trained knowledge. **Right:** The model outputs structured reactions, directly using these identifiers as “handles” to concisely represent complex molecules.

thermore, models supervised fine-tuned using the IdtVP strategy surpass those using BIVP, while also exhibiting stronger generalization.

Regarding the second challenge, current RxnDP fine-tuning relies entirely on Supervised Fine-Tuning (SFT) [3, 18, 22], which leads to two key limitations. First, **token-level objectives misalign with final evaluation metrics**. SFT provides token-wise supervision signals, whereas the task requires structured JSON outputs. Second, **the forced serialization of set-based tasks introduces noise**. Since RxnDP is inherently order-agnostic, SFT compels models to learn dataset-specific sequence orderings, consequently leading to overfitting, as shown in §Sec. 3.3.

To resolve this issue, we introduce the **Re³-DAPO** algorithm, employing verifiable reinforcement learning. Its core innovation lies in parsing the model-generated JSON outputs during reward computation and directly translating evaluation metrics into dense reward signals. This elevates the optimization objective from the token level to the reaction level, simultaneously eliminating serialized supervision signals. Experiments reveal that this method achieves significant performance gains over SFT baselines and proves effective under both BIVP and IdtVP strategies.

To evaluate model generalization, we introduce **ScannedRxn**, a unique benchmark dataset sourced from printed literature spanning the 1950s to the 1990s. It features authentic scanning artifacts and era-specific typography, diverging significantly from modern digital standards. This fills a critical gap present in existing natively-digital datasets, providing a rigorous testbed for assessing model robustness during the digitization of historical literature.

In summary, the main contributions of this paper are:

- **IdtVP Strategy:** We propose the “Identifier as Visual Prompt(IdtVP)” strategy, aligning with VLM pre-training patterns to activate built-in chemical knowledge and achieve robust zero-shot generalization.
- **Re³-DAPO Algorithm:** We introduce the Re³-DAPO algorithm, combining RLVR with mixed reward signals to address the misalignment of token-level objectives and serialization noise in set prediction tasks. Comprehensive experiments validate its significant performance improvements.
- **ScannedRxn Benchmark:** We release ScannedRxn, a highly challenging benchmark dataset containing printed literature from the 1950s to 1990s alongside authentic scanning artifacts, offering a rigorous testbed for evaluating the robustness of legacy chemical knowledge digitization.

2 Identifier as Visual Prompt strategy

2.1 Reaction Diagram Parsing Task

We formulate the Chemical Reaction Diagram Parsing (RxnDP) task as follows. Given an input diagram I , the RxnDP model $\mathcal{M}$ extracts a structured representation of all contained reactions, denoted as the set $\mathcal{R} = \{r_k\}$. This relationship is expressed as: $\mathcal{R} = \mathcal{M}(I)$. Each individual reaction $r_k \in \mathcal{R}$ is defined as a triple encompassing three key roles: $r_k = (\mathcal{S}_{\text{reactants}}, \mathcal{S}_{\text{conditions}}, \mathcal{S}_{\text{products}})$. For every role $\mathcal{S}_{\text{role}}$ (where $\text{role} \in \{\text{reactants}, \text{conditions}, \text{products}\}$), its contents form a set of components: $\mathcal{S}_{\text{role}} = \{e_m\}$. These components belong to one of two modalities: **1. Molecular Components:** Represented by the coordinates of the bounding box surrounding the corresponding molecular structure in the image I . **2. Textual Components:** Represented directly by their string-based textual description.

We adopt the reaction-level evaluation protocol from RxnCaption [22], reporting Precision, Recall, and F1 scores. A predicted reaction R^{pred} is a True Positive (TP) if it matches a ground truth R^{gt} based on the following metrics: **Soft Match:** Ignores textual content and merges conditions into reactants. A match requires an Intersection over Union (IoU) > 0.5 for all molecular bounding boxes. **Hybrid Match:** Tailored to output formats. For *BROS*: Adheres to the RxnScribe Hard Match standard, requiring IoU > 0.5 for all components (reactants, products, and conditions). For *BIVP/IdtVP*: Decouples structure and text. Molecules require IoU > 0.5 , while textual components are validated via string similarity, requiring a normalized edit distance ≤ 0.2 against the corresponding ground truth text.

2.2 Our Insights and IdtVP Strategy

For RxnDP, prior methods use **BROS** (**B**box and **R**ole in **O**ne **S**tep) to directly regress molecular coordinates. RxnCaption [22] argued this is ill-suited for VLMs and proposed **BIVP** (**B**ox **I**ndexing as **V**isual **P**rompts): detect molecules via MolYOLO, then overlay indexed boxes for the model to predict indices. BIVP

Table 1: Comparison of different visual prompting strategies under zero-shot setting. H-F1 and S-F1 denote Hybrid-F1 and Soft-F1 scores, respectively (%).

Strategy	RxnScribe-test		RxnCaption-15k-test	
	H-F1	S-F1	H-F1	S-F1
Gemini 3 Pro + BROS	3.90	71.20	4.92	49.80
Gemini 3 Pro + BIVP	45.27	81.32	47.40	73.40
Gemini 3 Pro + IdtVP	71.70	88.11	58.30	78.23
Qwen2.5-VL-7B + IdtVP	16.69	39.00	4.88	19.16
Intern-S1 + IdtVP	44.71	60.84	22.69	36.45

achieved strong performance (MolYOLO >98% precision/recall) with faster inference, providing crucial inspiration. However, pre-drawing boxes creates a distribution shift from VLM pretraining’s raw-image paradigm, inhibiting full activation of pretrained chemical knowledge.

We observe that, as illustrated in Fig. 1, chemical literature (particularly reaction diagrams) frequently annotates molecular structures with boldface numerals as identifiers (termed “molecule identifiers”). Such identifiers are not only ubiquitous within diagrams but also pervasive in accompanying paper text. Indeed, when describing reactions, chemical papers typically avoid cumbersome IUPAC names or SMILES strings, instead preferring concise identifiers to form expressions like “Using reactants **1** and **2** as starting materials, under catalysis by compound **3**, to afford product **4**.” In VLM pretraining corpora, such identifier-rich text often appears paired or interleaved with reaction diagrams, naturally conditioning models to adopt these identifiers as the preferred “language” for describing chemical reactions.

Building upon this insight, we propose the **Identifier as Visual Prompting (IdtVP)** strategy for RxnDP, with the workflow depicted in Fig. 3: (1) For molecules lacking identifiers in the chemical reaction diagram, we manually draw a unique, non-conflicting numerical identifier at an appropriate position adjacent to each molecule, ensuring every molecular structure is assigned a distinct identifier. (2) The model is then prompted to output chemical reactions using these molecular identifiers directly as “handles” to refer to corresponding molecules, as shown in Fig. 1.

2.3 Preliminary Verification of IdtVP Strategy

Zero-Shot Evaluation To validate the hypothesis that identifiers serve as visual prompt to activate the inherent chemical domain priors within VLMs, we conducted a preliminary zero-shot experiment comparing three strategies: BROS, BIVP, and IdtVP using representative models like Gemini 3 Pro [7]. As shown in Tab. 1, the BROS strategy (raw image input) failed to generate valid graph structures. While the BIVP strategy (bounding box prompting) provided spatial guidance, its performance remained limited, achieving only 45.27% Hybrid Match F1 on the RxnScribe-test dataset.

In contrast, the IdtVP strategy demonstrated significant superiority, boosting Gemini 3 Pro’s performance to 71.70% Hybrid Match F1 (and 88.11% Soft

Match F1) on RxnScribe-test, and maintaining a leading position on the more challenging RxnCaption-15k-test (78.23% Soft Match F1). These results strongly suggest that explicitly overlaid identifiers effectively bridge the semantic gap between visual features and reaction logic, thereby unlocking the model’s latent domain knowledge. Furthermore, the performance drop observed when applying IdtVP to a smaller parameter model (Qwen2.5-VL-7B [2], Intern-S1 [1]) indicates that high-precision zero-shot parsing requires a synergy of effective visual guidance and robust reasoning capabilities in the foundation model.

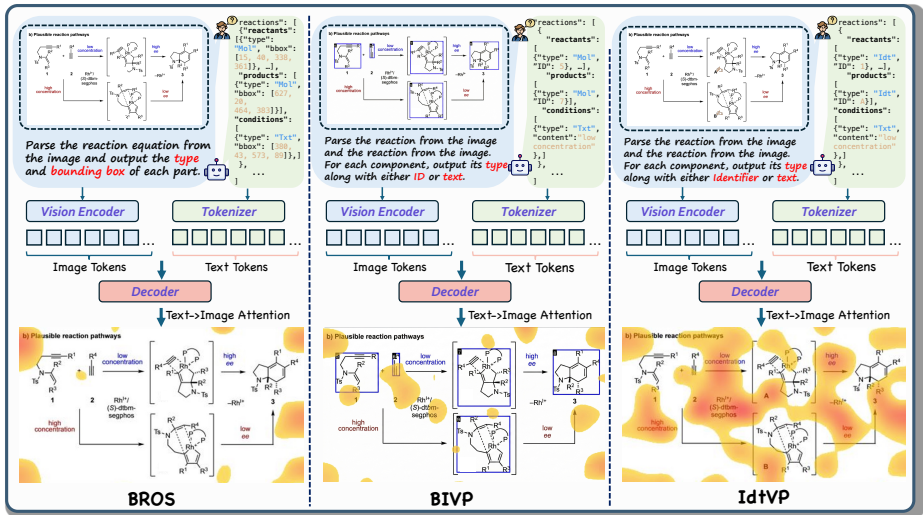


Fig. 2: Attention heatmaps of different prompting strategies. The top panels illustrate the input images, textual prompts, and corresponding output formats for each strategy. The bottom panels visualize the Text-to-Image attention heatmaps, computed by averaging the attention weights across all heads per layer, followed by average pooling across all layers. Notably, IdtVP yields far more precise and comprehensive visual grounding on molecules and text compared to BROS and BIVP.

Visual Attention Analysis To further validate our hypothesis, we adopt Layer-wise Attention Pooling [17] to visualize the cross-modal attention heatmaps of Qwen2.5-VL-7B [2] during the response generation stage, which reveal the visual features the model relies on when producing molecular or establishing referential relationships.

As illustrated in Fig. 2, the BROS strategy exhibits dispersed attention distributions, suggesting that the model struggles to establish strong semantic connections between the coordinate regression task and concrete chemical structures. The BIVP approach introduces explicit visual prompt in the form of bounding boxes to guide attention toward target regions. Although this leads to more local-

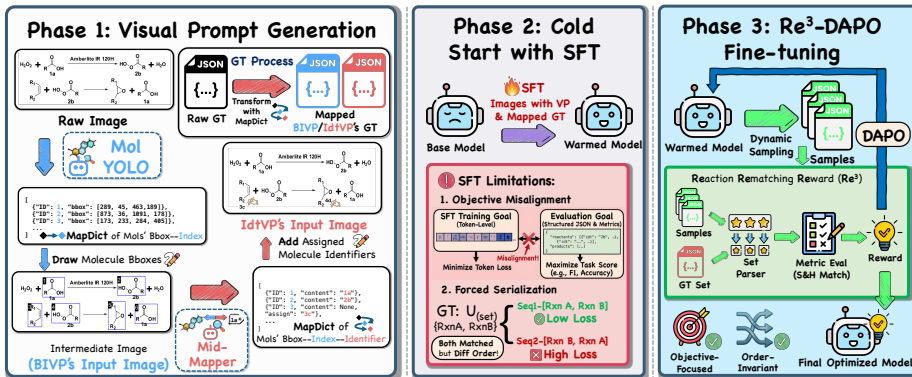


Fig. 3: Overview of the proposed framework. Phase 1 constructs IdtVP data effectively supporting both **zero-shot inference** and **model training**. Phases 2-3 introduce a generalized optimization paradigm (SFT followed by Re³-DAPO) that is transferable to other prompting variants (e.g., BIVP).

ized attention, the activations are frequently concentrated around the bounding box boundaries rather than the corresponding referential index. In contrast, the IdtVP strategy produces more concentrated and precise attention patterns, with activations consistently aligned with the identifiers and their molecular structures. This observation provides evidence for our hypothesis that VLMs may already encode latent representations that associate “identifiers” (e.g., 1a, 2b) with molecular entities from their pretraining. By constructing visual context that is more consistent with such representations, IdtVP appears to facilitate the activation of this latent knowledge without additional fine-tuning, leading to improved domain behavior compared to alternative strategies.

3 Method

3.1 Overall Framework

As illustrated in Fig. 3, the proposed framework parses chemical reaction diagrams via a progressive three-stage pipeline: Visual Prompt Generation, Cold Start with SFT and Re³-DAPO Fine-tuning.

In Phase 1, we utilize our Mid-Mapper (§Sec. 3.2) to construct the IdtVP by explicitly mapping identifiers to molecular bounding boxes. This format supports zero-shot inference and serves as a high-quality foundation for model training.

Phase 2 establishes the model’s initial capabilities via SFT. While this serves as a necessary cold start, it inevitably inherits the optimization bottlenecks discussed in §Sec. 1 (i.e., objective misalignment and permutation sensitivity). To overcome these limitations, Phase 3 introduces the Re³-DAPO fine-tuning strategy (§Sec. 3.3). By shifting the learning objective from **next-token prediction** to **set-level matching**, we align the optimization directly with the evaluation metrics. Notably, this training curriculum (Phase 2 → Phase 3) constitutes a

generalized optimization paradigm, applicable not only to IdtVP but also transferable to other prompting variants (e.g., BIVP).

3.2 Mid-Mapper: Molecule-Identifier Mapper

To establish a semantic association between chemical identifiers and molecular structures in reaction diagrams, we propose **Molecule-Identifier Mapper (Mid-Mapper)**. Leveraging the precise molecular localization provided by MolYOLO, the Mid-Mapper enables accurate alignment between each detected molecular structure and its corresponding chemical identifier.

Specifically, we distilled Gemini 3 Pro to construct high-quality training data that maps visual localization signals to semantic identifiers. For molecules that lack explicit identifiers in the original reaction diagrams, we required Gemini to assign context-consistent virtual identifiers based on reaction mechanism, ensuring that every molecular structure localized by MolYOLO has a unique and non-overlapping identifier. Building upon this dataset, we performed full-parameter fine-tuning of Qwen2.5-VL-3B to learn robust molecule–identifier associations. At inference time, the predicted virtual identifiers are plotted back into the original diagrams, thereby providing complete and semantically aligned visual annotations. This design ensures that the downstream IdtVP framework can perform efficient reasoning based on comprehensive and well-grounded visual features. The details of Mid-Mapper are provided in the Supplementary Material.

3.3 Re³-DAPO Training

Standard Supervised Fine-Tuning (SFT) relies on token-level teacher-forcing, making it highly sensitive to the serialization order of target data. This is particularly problematic for the RxnDP task, where chemical reaction diagrams depict multiple independent reactions and entities that are inherently permutation-invariant. To quantify this limitation, we evaluated fully converged SFT models (BIVP and IdtVP) on the training set (see §Sec. 4.1 for details on dataset construction), explicitly isolating samples where the models predicted chemically perfect structures. As shown in Tab. 2, even on these semantically accurate samples, both models incurred considerable cross-entropy loss and generated non-zero gradients solely due to formatting mismatches with the ground truth.

Table 2: Analysis of sequential inconsistency in SFT models. Image-level statistics exclude single-reaction samples.

Model	Image-level			Reaction-level		
	Total	Errors	Rate (%)	Total	Errors	Rate (%)
BIVP	3,882	500	12.88	20,992	3,131	14.92
IdtVP	5,833	272	4.66	31,300	2,538	8.11

This confirms that rigid teacher-forcing compels the network to memorize arbitrary data formats rather than genuine chemical semantics. To address this disconnect between token-level optimization and order-agnostic, reaction-level accuracy, we propose the Re³-DAPO algorithm. Re³-DAPO leverages verifiable reinforcement learning to compute rewards based purely on the structured chemical entities, effectively bypassing the noise of arbitrary serialization.

Specifically, we define a composite, permutation-invariant reward function $R(y)$ that translates discrete evaluation metrics into a dense signal. As introduced in §Sec. 2.1, it averages Hybrid Match (R_{hybrid}) and Soft Match (R_{soft}) to prioritize molecular accuracy over rigid role categorization:

$$R(y) = \frac{1}{2} (R_{\text{hybrid}}(y, y_{gt}) + R_{\text{soft}}(y, y_{gt})) \quad (1)$$

We employ DAPO to directly maximize this expected reward. By elevating the optimization from the token level to the reaction level, Re³-DAPO allows the model to explore valid JSON serializations anchored solely by an order-invariant semantic objective, eliminating the need for complex actor-critic architectures.

3.4 ScannedRxn Dataset

Existing chemical reaction parsing benchmarks, such as RxnCaption-15k-test [22] and RxnScribe-test [18], are primarily derived from contemporary literature and lack the printed data styles common in the 20th century. Consequently, they fail to fully evaluate model generalization on historical archives, which is crucial for comprehensively extracting chemical knowledge and preserving valuable scientific discoveries. To bridge this gap, we introduce the ScannedRxn dataset. Collected from publications spanning the 1950s to the 1990s, this dataset comprises 200 reaction diagrams evenly distributed across four types: single, multi-line, tree, and cyclic (50 samples each). It effectively captures the distribution of traditional printed styles, including scanning noise, typewriter fonts, and legacy layout conventions. Comprehensive construction details and illustrative examples of ScannedRxn are provided in the Supplementary Material.

4 Experiment

4.1 Experiment Setup

Dataset Construction To construct the training dataset for IdtVP (denoted as RxnID), we built upon the official open-source RxnCaption-15k dataset. Specifically, we utilized the trained Mid-Mapper to detect existing identifiers within these reaction diagrams. For molecules lacking explicit identifiers, we assigned new ones and employed automated tools to directly draw them onto the original images. Both the resulting dataset configuration and the adopted data augmentation strategies strictly follow the RxnCaption [22] setting. Meanwhile, the SFT and RL (Re³-DAPO) training of BIVP were directly performed on the unmodified RxnCaption-15k dataset.

Table 3: Model performance comparison on RxnScribe-test, RxnCaption-15k-test, and the new ScannedRxn datasets. Best scores are in **bold**, second best are underlined. +RL means post trained by **Re³-DAPO**, same as below.

Model	Strategy	RxnScribe-test						RxnCaption-15k-test						ScannedRxn					
		Hybrid Match			Soft Match			Hybrid Match			Soft Match			Hybrid Match			Soft Match		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Trained Models																			
RxnID+RL	IdtVP	76.3	73.7	75.0	87.2	84.6	85.9	65.5	63.4	64.4	75.8	73.2	74.5	60.2	52.8	56.3	82.5	70.8	76.2
RxnID	IdtVP	75.8	73.5	74.6	87.2	84.5	85.6	64.2	61.1	61.2	75.2	70.6	72.8	58.1	51.4	54.5	80.5	69.1	74.4
RxnCaption+RL	BIVP	76.0	75.8	75.9	86.8	87.1	86.9	65.5	62.9	64.1	75.3	72.0	73.6	54.2	38.6	45.1	72.2	50.2	59.2
RxnCaption	BIVP	71.6	72.7	72.2	85.3	87.1	86.2	60.3	59.3	59.8	71.3	69.4	70.4	54.6	47.9	51.0	73.1	66.2	69.5
	BROS	69.6	68.9	69.2	76.2	76.2	76.2	57.0	57.5	57.2	66.4	67.4	66.9	56.3	52.3	54.2	68.7	67.6	68.2
RxnScribe_w/15k	BROS	72.4	69.1	70.7	84.1	81.7	82.8	61.2	38.7	47.4	72.1	44.7	55.2	60.7	43.3	50.5	76.1	58.2	65.9
RxnScribe	BROS	72.3	66.2	69.1	83.8	76.5	80.0	47.4	27.6	34.9	62.1	36.4	45.9	57.6	42.8	49.1	75.5	59.7	66.7
RxnIM	BROS	71.0	70.1	70.5	79.2	74.7	76.9	48.8	30.3	37.4	52.9	32.8	40.5	43.1	15.5	22.8	48.6	23.4	31.6
Open-source Models																			
Intern-S1	BIVP	49.9	53.1	51.4	66.4	71.2	68.8	27.8	24.4	26.0	44.7	40.2	42.3	11.2	9.3	10.2	23.7	19.9	21.7
	IdtVP	53.8	38.3	44.7	72.7	52.3	60.8	33.8	17.1	22.7	53.3	27.7	36.5	44.5	35.5	39.5	71.4	58.4	64.3
Qwen2.5-VL-7B	BIVP	6.0	4.1	4.9	55.8	36.0	43.8	2.9	0.9	1.4	33.0	10.3	15.6	37.1	31.7	34.2	51.3	45.3	48.1
	IdtVP	21.2	13.8	16.7	49.0	32.4	39.0	10.1	3.2	4.9	38.6	12.7	19.2	40.9	30.1	34.7	63.5	50.6	56.3
Closed-source Models																			
Gemini 3.0 Pro	BIVP	41.7	49.5	45.3	74.4	89.6	81.3	46.5	48.4	47.4	71.6	75.3	73.4	18.3	17.8	18.6	51.2	49.5	50.3
	IdtVP	72.0	71.4	71.7	87.9	88.3	88.1	59.0	57.6	58.3	78.5	78.0	78.2	60.7	58.0	59.3	89.4	83.9	86.5
GPT-4o	BIVP	19.7	25.3	22.1	47.9	59.1	52.9	14.0	16.0	15.0	27.5	29.9	28.7	10.9	8.9	9.8	27.9	19.3	22.8
	IdtVP	37.7	37.8	37.7	66.3	67.4	66.8	20.9	17.3	18.3	48.8	36.7	41.9	41.8	30.6	35.3	65.8	52.2	58.2
Qwen-VL-Max	BIVP	43.1	44.4	43.7	70.1	73.3	71.7	34.7	26.2	29.8	63.6	48.8	55.2	21.3	16.3	18.5	34.6	27.8	30.8
	IdtVP	48.7	42.9	45.6	72.4	64.5	68.2	26.8	25.3	26.0	44.9	43.1	44.0	49.7	36.2	41.9	73.9	59.4	65.9

Mid-Mapper Mid-Mapper was developed by performing full-parameter SFT on Qwen2.5-VL-3B via the ms-swift framework. Training was conducted on two NVIDIA A100 GPUs for 1 epoch with a learning rate of 1×10^{-5} , a 5% warmup ratio, and a per-device batch size of 2.

RxnDP Model The training of RxnCaption and RxnID consisted of two stages:

- **Cold-Start SFT:** To ensure schema adherence, we performed an SFT cold-start using approximately 10% of the training data, deliberately selecting the checkpoint where the loss curve began to plateau. This brief SFT prevents model memorization while providing a robust initialization for JSON schema constraints. We fine-tuned the model using ms-swift on 8 NVIDIA A100 GPUs. We utilized the AdamW optimizer with a maximum learning rate of 1×10^{-5} , a cosine decay scheduler, and a 5% warmup ratio. To handle long contexts, we enabled DeepSpeed ZeRO Stage 2 and set the per-device batch size to 1 with 16 gradient accumulation steps, resulting in an effective batch size of 128.
- **Re³-DAPO Fine-tuning:** Subsequently, we applied our proposed DAPO algorithm (implemented via the VeRL framework) on a single compute node equipped with 8 NVIDIA H200 GPUs. The model was optimized using AdamW with a learning rate of 1×10^{-6} and a global batch size of 256,

utilizing FSDP for model parallelism. To stabilize training and reduce variance, we adopted the ‘‘Clip-Higher’’ strategy with a DAPO clip range of $[0.2, 0.28]$, a KL penalty coefficient of 0.01, and a rollout size of $N = 8$.

Thanks to the SFT cold-start, format violations were rare; however, if the model generated unparseable or invalid JSON during RL exploration, we assigned a strict accuracy reward of 0, effectively teaching the model to maintain valid syntax autonomously.

The total RL training consisted of 870 steps (approximately 5 epochs on the full dataset). Due to the substantial computational cost of RL training on the VLM (requiring approximately 70 hours on a node with $8 \times$ NVIDIA H200 GPUs), we were restricted to a single full-scale training run. Therefore, we did not perform multi-run statistical variance analysis.

Baselines We evaluated the zero-shot performance of current mainstream general VLMs using identifiers as visual prompt, including GPT-4o [12], Gemini 3.0 Pro [7](with all metrics assessed in February 2026), Qwen2.5-VL-7B [2], and Intern-S1 [1]. Additionally, we tested specialized models for the RxnDP task, primarily RxnScribe [18], RxnIM [3], and the BIVP version of RxnCaption-VL [22].

4.2 Main Results

Strategies and Prior Knowledge The main results are shown in Tab. 3. The choice of visual prompt strategy is the primary determinant of model performance. As shown in Table 4, IdtVP consistently demonstrates a superior advantage over BIVP and BROS. Notably, IdtVP effectively activates the latent chemical reasoning and structural prior knowledge of large-scale foundation models like Gemini 3.0 Pro. For instance, Gemini 3.0 Pro (IdtVP) achieves a state-of-the-art (SOTA) Soft Match F1 of 86.5% on the historical ScannedRxn dataset, even in a zero-shot setting. While BROS struggles with complex, real-world layouts and BIVP provides only coarse spatial cues, IdtVP offers fine-grained identifier-level guidance. Our RxnID+RL further leverages this strategy to reach a 64.4% Hybrid Match F1 on RxnCaption-15k-test, surpassing all open-source baselines and competitive closed-source models such as GPT-4o [12].

Re³-DAPO Performance and Stability While Re³-DAPO significantly improves in-distribution precision when applied to both BIVP and IdtVP strategies, the out-of-distribution stability of the resulting models varies sharply. To emphasize the impact of the underlying parsing strategies, we hereafter denote RxnCaption with Re³-DAPO as BIVP+RL, and RxnID with Re³-DAPO as IdtVP+RL. On the RxnCaption-15k-test set, both BIVP+RL and IdtVP+RL exhibit marked performance gains. However, when facing the severe domain shift of ScannedRxn, BIVP+RL often collapses into degenerate token repetition, leading to parsing failures and performance drops. Conversely, IdtVP+RL maintains high stability and avoids such looping hallucinations; the robust structural anchor provided by IdtVP facilitates reinforcement learning optimization, achieving a Hybrid Match F1 of 56.3% on scanned data.

Besides, it is notably that RL nearly eliminates the performance gap between BIVP and IdtVP on in-distribution test sets. This directly proves our analysis in §Sec. 3.3 and Tab. 2 regarding **serialization noise**. While IdtVP naturally incurs fewer image-level and reaction-level serialization errors caused by token-level optimized SFT, reaction-level RL optimization effectively neutralizes this discrepancy, allowing BIVP to overcome its initial serialization penalty.

Generalization and Scale The evaluation on RxnScribe-test reveals limited gains (e.g., 75.9% Hybrid Match F1 for RxnCaption+RL), likely due to its small sample size (138 images) and inherent annotation inconsistencies, where minor score variations lack statistical significance. In contrast, the larger RxnCaption-15k-test set and the historical ScannedRxn dataset provide a more rigorous assessment of generalization. The latter highlights the severe limitations of synthetic-only training, as seen in RxnIM’s low 22.8% Hybrid-F1, whereas our RL-refined models demonstrate the ability to capture universal reaction logic across divergent document eras.

4.3 Ablation and Analysis

Influence of Identifier Detector Tab. 4 demonstrates that Mid-Mapper, as a lightweight model distilled from Gemini, achieves performance **comparable** to its teacher model. For instance, when evaluating with IdtVP+RL, the Soft Match F1 gap between Mid-Mapper and teacher model is merely 1.1%, confirming the effectiveness of our distillation. Interestingly, using human-annotated ground truth as detector leads to slightly lower performance than using either Gemini or Mid-Mapper. Since IdtVP+RL was trained on Gemini-generated data, the natural inconsistencies in human labeling style and formatting essentially act as noise. Mid-Mapper avoids these by inheriting the teacher’s stable output patterns, resulting in better compatibility with downstream tasks.

Table 4: Performance comparison of different identifier detectors on the RxnCaption-15k-test set. We evaluate the impact of detector precision on the Soft-F1 and Hybrid-F1 scores across different VLM prompting strategies.

Detector	VLM	RxnCaption-15k-test	
		Soft-F1	Hybrid-F1
Mid-Mapper	IdtVP+RL	74.5	64.4
	Gemini	78.2	58.3
Gemini	IdtVP+RL	75.6	64.7
	Gemini	78.6	60.2
Ground Truth	IdtVP+RL	73.6	63.1
	Gemini	77.9	58.0

Table 5: Main results on test set (RS=RxnScribe, RC=RxnCaption-15k-test, SR=ScannedRxn). Met. and Drp.(↓) denote the Metric value and the performance drop relative to the upper bound, respectively.

Error Source	RS		RC		SR	
	Met.	Drp↓	Met.	Drp↓	Met.	Drp↓
<i>Upper Bound</i>	100.0	-	100.0	-	100.0	-
MolYOLO Err. (Ideal RxnDP VLM)	99.7	0.3	95.3	4.7	87.1	12.9
Mid-Mapper Err. (Ideal RxnDP VLM)	98.5	1.2	93.8	1.5	84.7	2.4
Assign Error (Soft-F1)	85.9	12.6	74.5	19.3	76.2	8.5
Textual Error (Hybrid-F1)	75.0	10.9	64.4	10.1	56.3	19.9

Stage-wise Error Analysis To isolate performance bottlenecks, we perform a stage-wise error analysis on our top-performing RxnID+RL model (Tab. 5). **MolYOLO Err.** and **Mid-Mapper Err.** represent the upper-bound loss from upstream detectors during visual prompt generation, assuming an ideal RxnDP VLM. **Assign Error** (Soft-F1) and **Textual Error** (Hybrid-F1) denote losses from incorrect index assignments and OCR failures, respectively.

Quantitative results show that **early-stage detectors contribute minimally** to overall error. However, on blurrier ScannedRxn, both MolYOLO and OCR become significant bottlenecks. Otherwise, RxnDP VLM remains the main bottleneck, struggling with complex graph topologies (e.g., dense cyclic flows) and unconventional visual symbols (e.g., lightbulb or lightning icons). Even with the cascading losses typical of multi-stage designs, **RxnID still surpass the trained, end-to-end BROS model** by roughly 10% on all three datasets. Fine-grained metrics and error examples are provided in the Supplementary Material.

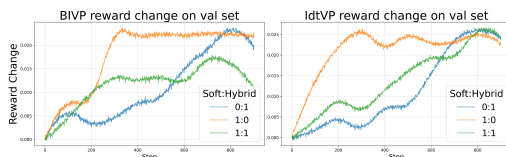


Fig. 4: Reward Δ change on the val dataset. The Y-axis denotes the reward improvement relative to the initial step (Step 0).

Table 6: Quantitative results on RxnCaption-15k-test. Comparison across reward ratios.

Ratio	Strategy	RxnCaption-15k-test	
		Soft-F1	Hybrid-F1
0 : 1	BIVP+RL	72.5	64.5
	IdtVP+RL	73.6	<u>64.4</u>
1 : 0	BIVP+RL	73.8	62.7
	IdtVP+RL	<u>74.2</u>	62.8
1 : 1	BIVP+RL	73.6	64.1
	IdtVP+RL	74.5	<u>64.4</u>

Influence of Reward Function We ablated Soft:Hybrid reward ratios (1:0, 0:1, 1:1), monitoring reward Δ change on the val dataset and evaluating final performance on RxnCaption-15k-test.

- **Training Dynamics:** As shown in Fig. 4, the Soft-dominant (1:0) setting facilitates rapid initial convergence but plateaus early due to relaxed criteria. In contrast, the Hybrid-dominant (0:1) setting shows a slower start but maintains stronger upward momentum in later stages.
- **Final Performance:** As show in Tab. 6, the balanced (1:1) configuration achieves the most stable trajectory and best generalization on RxnCaption-15k-test. It effectively combines the early-stage guidance of Soft Match with the precise optimization constraints of Hybrid Match, leading to the best overall performance across both models.

4.4 Reinforcement Learning Stability and Significance

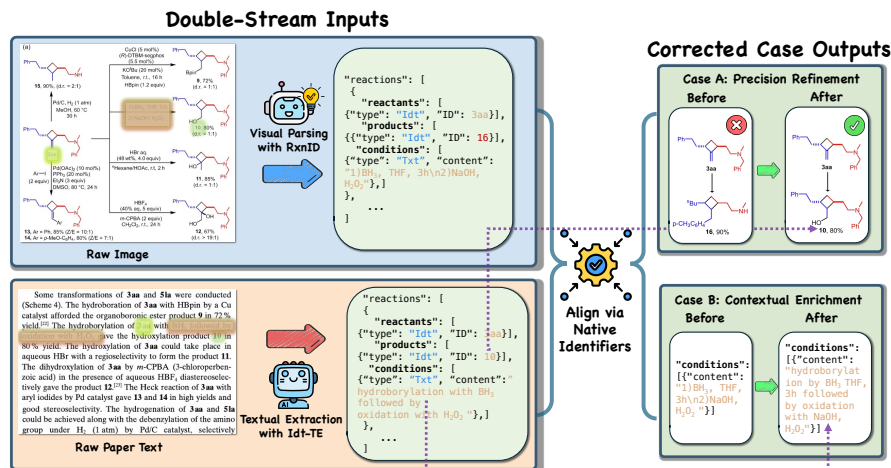
Since repeating full 7B reinforcement learning experiment is expensive, we added a 3B multi-seed study ($4 \times$ H200 respectively, ~ 2 days) to test whether Re³-DAPO gains are stable across seeds and consistent with the 7B trend. Starting

Table 7: 3B multi-seed Re³-DAPO results on Hybrid F1. We report results over 5 RL seeds. All rows improve in 5/5 seeds with exact one-sided sign-test $p = 0.03125$.

Setting	Dataset	Cold RL	mean $\pm$ std	95% CI	Δ
BIVP	RxnCaption-15k-test	13.37	54.57 $\pm$ 3.58	3.13	+41.20
BIVP	RxnScribe-test	57.11	57.93 $\pm$ 0.63	0.56	+0.82
BIVP	ScannedRxn	5.79	7.03 $\pm$ 0.61	0.53	+1.24
IdtVP	RxnCaption-15k-test	30.39	57.90 $\pm$ 0.35	0.31	+27.51
IdtVP	RxnScribe-test	68.52	73.51 $\pm$ 0.68	0.59	+4.99
IdtVP	ScannedRxn	39.56	57.74 $\pm$ 0.74	0.65	+18.18

from the same cold-start checkpoint, we ran 5 seeds for both BIVP and IdtVP and evaluated the best checkpoint on three official test sets. Once the cold-start model reliably parses valid JSON, the exact plateau point has limited impact on the best RL checkpoint. Tab. 7 shows that RL improves Hybrid F1 over cold-start in all 5/5 seeds, with exact one-sided sign-test $p = 0.03125$. The seed variance is generally much smaller than the mean gain, suggesting that the improvement is stable across seeds rather than a lucky run. The gain on RxnScribe is smaller (RxnID vs. RxnID+RL 7B, +0.4%) because its layouts are relatively simple, leaving mainly corner cases and less room for improvement.

4.5 IdtVP’s Exclusive Cross-Modal Extensibility

**Fig. 5: Overview of the Cross-Modal Verification pipeline.** System processes double-stream inputs: visual parsing of raw diagrams via RxnID (top) and textual extraction from the manuscript via Idt-TE (bottom), enabling downstream applications such as **Precision Refinement** (Case A) and **Contextual Enrichment** (Case B).

Unlike BIVP or Set-of-Mark which rely on isolated geometric anchors, IdtVP establishes a robust semantic bridge between vision and language. By adopting the **author’s native identifiers** (e.g., **1a**) prevalent in both diagrams and text, IdtVP uniquely enables unified cross-modal indexing.

To demonstrate this exclusive advantage, we designed a **Cross-Modal Verification** pipeline (Fig. 5), which processes double-stream inputs via visual parsing (**RxnID**) and textual extraction (**Idt-TE**). By leveraging this shared vocabulary, the pipeline aligns LLM-extracted textual data with visual predictions seamlessly, without requiring any intermediate mapping. This exclusive capability enhances system robustness through:

- **Precision Refinement (Fig. 5, Case A)**: Resolving visual ambiguities or OCR errors using textual context. This cross-referencing is inherently impossible for BIVP’s arbitrary indices (e.g., Box #1), which lack any semantic correlate in the manuscript text.
- **Contextual Enrichment (Fig. 5, Case B)**: Completing visual graphs with hidden attributes retrieved from the text. This positions IdtVP not merely as a parsing tool, but as the foundational bridge for a truly unified multimodal chemical knowledge extraction system.

Admittedly, while our Idt-TE pipeline comprehensively extracts rich reaction data from text, optimally applying these insights to the RxnDP task remains an open area for exploration, such as handling low-confidence refinements and determining if contextual enrichment exceeds the strict scope of pure diagram parsing. While exploring these integration strategies is left for future work, IdtVP’s native identifiers uniquely provide the indispensable foundation for such unified extraction. Detailed Idt-TE’s pipeline architectures, exact LLM prompts, each stage’s output and representative qualitative case studies are provided in the Supplementary Material.

5 Conclusion

In this paper, we propose RxnID, a framework leveraging identifiers as visual prompts to activate VLM domain priors for precise chemical reaction parsing. To overcome optimization misalignment and forced serialization noise in SFT, we introduce Re³-DAPO, a reinforcement learning strategy with permutation-invariant rewards that enforces sequence-level correctness. Beyond visual parsing, IdtVP establishes a semantic bridge for Cross-Modal Verification, enabling precision refinement and contextual enrichment. Experiments on RxnCaption-15k, RxnScribe, and ScannedRxn confirm that RxnID and Re³-DAPO outperform prior approaches, offering a robust and extensible solution for digitizing chemical archives.

Acknowledgements

This work was supported by Shanghai Artificial Intelligence Laboratory.

References

1. Bai, L., Cai, Z., Cao, Y., Cao, M., Cao, W., Chen, C., Chen, H., Chen, K., Chen, P., Chen, Y., et al.: Intern-s1: A scientific multimodal foundation model. arXiv preprint arXiv:2508.15763 (2025)
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
3. Chen, Y., Leung, C.T., Sun, J., Huang, Y., Li, L., Chen, H., Gao, H.: Towards large-scale chemical reaction image parsing via a multimodal large language model. arXiv preprint arXiv:2503.08156 (2025)
4. Ding, H., Hua, P., Huang, Z.: Survey on recent progress of ai for chemistry: Methods, applications, and opportunities. arXiv preprint arXiv:2502.17456 (2025)
5. Fan, V., Qian, Y., Wang, A., Wang, A., Coley, C.W., Barzilay, R.: Openchemie: An information extraction toolkit for chemistry literature. *Journal of Chemical Information and Modeling* **64**(14), 5521–5534 (2024)
6. Gao, J., Xu, S., Ye, W., Liu, W., He, C., Fu, W., Mei, Z., Wang, G., Wu, Y.: On designing effective rl reward at training time for llm reasoning (2024), <https://arxiv.org/abs/2410.15115>, accessed: 2026-06-30
7. Google Gemini Team: Gemini 3 pro: A multimodal large language model (2026), <https://gemini.google.com/>, accessed: 2026-03-04
8. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
9. Guo, J., Ibanez-Lopez, A.S., Gao, H., Quach, V., Coley, C.W., Jensen, K.F., Barzilay, R.: Automated chemical reaction extraction from scientific literature. *Journal of chemical information and modeling* **62**(9), 2035–2045 (2021)
10. Hu, J., Zhang, Y., Han, Q., Jiang, D., Zhang, X., Shum, H.Y.: Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. arXiv preprint arXiv:2503.24290 (2025)
11. Huang, W., Jia, B., Zhai, Z., Cao, S., Ye, Z., Zhao, F., Xu, Z., Hu, Y., Lin, S.: Vision-r1: Incentivizing reasoning capability in multimodal large language models. arXiv preprint arXiv:2503.06749 (2025)
12. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
13. Lambert, N., Morrison, J., Pyatkin, V., Huang, S., Ivison, H., Brahman, F., Miranda, L.J.V., Liu, A., Dziri, N., Lyu, S., et al.: Tulu 3: Pushing frontiers in open language model post-training. arXiv preprint arXiv:2411.15124 (2024)
14. Laurençon, H., Saulnier, L., Tronchon, L., Bekman, S., Singh, A., Lozhkov, A., Wang, T., Karamcheti, S., Rush, A., Kiela, D., et al.: Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems* **36**, 71683–71702 (2023)
15. Li, Q., Chen, Z., Wang, W., Wang, W., Ye, S., Jin, Z., Chen, G., He, Y., Gao, Z., Cui, E., et al.: Omnicorpus: A unified multimodal corpus of 10 billion-level images interleaved with text. arXiv preprint arXiv:2406.08418 (2024)
16. Li, Y., Jiang, J., Zhu, J., Peng, S., Wei, B., Zhou, Y., Gao, L.: Uni-mumer: Unified multi-task fine-tuning of vision-language model for handwritten mathematical expression recognition. arXiv preprint arXiv:2505.23566 (2025)

17. Oh, D., Kim, Y., Lee, H., Huang, H.H., Lim, H.: Don't judge a language model by its last layer: Contrastive learning with layer-wise attention pooling. arXiv preprint arXiv:2209.05972 (2022)
18. Qian, Y., Guo, J., Tu, Z., Coley, C.W., Barzilay, R.: Rxnscribe: a sequence generation model for reaction diagram parsing. *Journal of chemical information and modeling* **63**(13), 4030–4041 (2023)
19. Schneider, N., Lowe, D.M., Sayle, R.A., Tarselli, M.A., Landrum, G.A.: Big data from pharmaceutical patents: a computational analysis of medicinal chemists' bread and butter. *Journal of medicinal chemistry* **59**(9), 4385–4402 (2016)
20. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. arXiv preprint arXiv:2402.03300 **2**(3), 5 (2024)
21. Shen, H., Liu, P., Li, J., Fang, C., Ma, Y., Liao, J., Shen, Q., Zhang, Z., Zhao, K., Zhang, Q., et al.: Vlm-r1: A stable and generalizable r1-style large vision-language model. arXiv preprint arXiv:2504.07615 (2025)
22. Song, J., Wang, C., Jiang, B., Wang, Y., Zheng, H., Wei, X., Liu, C., Gao, J., Wang, Y., Wu, L., et al.: Rxncaption: Reformulating reaction diagram parsing as visual prompt guided captioning. arXiv preprint arXiv:2511.02384 (2025)
23. Wang, P., Ma, R., Zhang, B., Chen, X., He, Z., Luo, K., Lv, Q., Jiang, Q., Xie, Z., Wang, S., et al.: Rlver: Reinforcement learning with verifiable emotion rewards for empathetic agents. arXiv preprint arXiv:2507.03112 (2025)
24. Wen, X., Liu, Z., Zheng, S., Ye, S., Wu, Z., Wang, Y., Xu, Z., Liang, X., Li, J., Miao, Z., et al.: Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms. arXiv preprint arXiv:2506.14245 (2025)
25. Wilary, D.M., Cole, J.M.: Reactiondataextractor 2.0: a deep learning approach for data extraction from chemical reaction schemes. *Journal of Chemical Information and Modeling* **63**(19), 6053–6067 (2023)
26. Yu, E., Lin, K., Zhao, L., Yin, J., Wei, Y., Peng, Y., Wei, H., Sun, J., Han, C., Ge, Z., et al.: Perception-r1: Pioneering perception policy with reinforcement learning. arXiv preprint arXiv:2504.07954 (2025)
27. Yu, Q., Zhang, Z., Zhu, R., Yuan, Y., Zuo, X., Yue, Y., Dai, W., Fan, T., Liu, G., Liu, L., et al.: Dapo: An open-source llm reinforcement learning system at scale. arXiv preprint arXiv:2503.14476 (2025)
28. Zhong, M., Ouyang, S., Jiang, M., Hu, V., Jiao, Y., Wang, X., Han, J.: Reactie: Enhancing chemical reaction extraction with weak supervision. arXiv preprint arXiv:2307.01448 (2023)
29. Zhong, M., Ouyang, S., Jiao, Y., Kargupta, P., Luo, L., Shen, Y., Zhou, B., Zhong, X., Liu, X., Li, H., et al.: Reaction miner: An integrated system for chemical reaction extraction from textual data. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. pp. 389–402 (2023)
30. Zhu, J., Zhao, W., Li, Y., Hu, X., Gao, L.: Tamer: tree-aware transformer for handwritten mathematical expression recognition. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. pp. 10950–10958 (2025)
31. Zhu, W., Hessel, J., Awadalla, A., Gadre, S.Y., Dodge, J., Fang, A., Yu, Y., Schmidt, L., Wang, W.Y., Choi, Y.: Multimodal c4: An open, billion-scale corpus of images interleaved with text. *Advances in Neural Information Processing Systems* **36**, 8958–8974 (2023)